CSE3506 Essentials of Data Analytics

Name : **Revanth Bhargava Boyidi**
Reg. No. : **19BCE1177**
Lab Exercise: **3-ANOVA**

Objective: To apply ANOVA on the analytical and experimental values in the dataset using R programming.

Methods:

i. Determine the grand mean, and mean of thermal conductivity.

ii. Sum of Squares between and Sum of Squares within

iii. Degrees of Freedom (Between, Within and total)

iv. Mean Squares (Between and Within)

v. F_Statistic and F Critical

vi. Check, Is Null hypothesize accepted?

vii. Plot Vol. Concentration VS Thermal Conductivity

viii. Plot Vol Concentration Vs % increase in Thermal conductivity –(In bar chart)

ix. Error (Measured Value-Analytical Value)

x. Conclusion

Setting the path and Importing libraries

```
setwd("D:\\SEM-VI\\EDA_CSE3506\\Lab\\Lab-3(28-01)_Anova")
rm(list=ls())      #To clear the environment
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library("ggplot2")

#Importing Dataset
data <- read.csv("Dataset.csv",header=TRUE)
data

##   Vol.Concentration Analytical_K X.Increase.in.K Experimental_K  Erro
r
## 1              0.00      0.21400          0.0000          0.214 0.0000
## 2              0.01      0.39186         83.1121          0.450 0.0581
## 3              0.02      0.56972        166.2243          0.650 0.0803
## 4              0.03      0.74758        249.3364          0.800 0.0524
## 5              0.04      0.92544        332.4486          1.000 0.0746
## 6              0.05      1.10330        415.5607          1.200 0.0967
## 7              0.10      1.99260        831.1215          2.100 0.1074
## 8              0.15      2.88190       1246.6822          3.000 0.1181
##   X.Error.in.K
## 1           NA
## 2    12.920000
## 3    12.350769
## 4     6.552500
## 5     7.456000
## 6     8.058333
## 7     5.114286
## 8     3.936667

data[is.na(data)] = 0 #Filling null values with 0
summary(data)

##  Vol.Concentration  Analytical_K    X.Increase.in.K  Experimental_K
##  Min.   :0.0000    Min.   :0.2140   Min.   :   0.0   Min.   :0.214
##  1st Qu.:0.0175    1st Qu.:0.5253   1st Qu.: 145.4   1st Qu.:0.600
##  Median :0.0350    Median :0.8365   Median : 290.9   Median :0.900
##  Mean   :0.0500    Mean   :1.1033   Mean   : 415.6   Mean   :1.177
##  3rd Qu.:0.0625    3rd Qu.:1.3256   3rd Qu.: 519.5   3rd Qu.:1.425
##  Max.   :0.1500    Max.   :2.8819   Max.   :1246.7   Max.   :3.000
##      Error           X.Error.in.K
##  Min.   :0.00000   Min.   : 0.000
##  1st Qu.:0.05668   1st Qu.: 4.820
##  Median :0.07745   Median : 7.004
##  Mean   :0.07345   Mean   : 7.049
##  3rd Qu.:0.09938   3rd Qu.: 9.131
##  Max.   :0.11810   Max.   :12.920
```

(i) Determine the grand mean, and mean of thermal conductivity.

The grand mean of a set of samples is the total of all the data values divided by the total sample size. This requires that you have all of the sample data available to you, which is usually the case, but not always. It turns out that all that is necessary to find perform a one-way analysis of variance are the number of samples, the sample means, the sample variances, and the sample sizes.

```
gm = mean(c(mean.default(data$Analytical_K), mean.default(data$Experime
ntal_K)))
print(paste("Grand Mean:", gm))

## [1] "Grand Mean: 1.140025"

am = mean(data$Analytical_K)
print(paste("Mean of Analytical Thermal Conductivity:", am))

## [1] "Mean of Analytical Thermal Conductivity: 1.1033"

em = mean(data$Experimental_K)
print(paste("Mean of Experimental Thermal Conductivity:", em))

## [1] "Mean of Experimental Thermal Conductivity: 1.17675"
```

ANOVA Table

ANOVA stands for Analysis of Variance. One-Way Analysis of Variance tells you if there are any statistical differences between the means of three or more independent groups. ANOVA is used to test a particular hypothesis. ANOVA helps to understand how different groups respond, with a null hypothesis for the test that the means of the different groups are equal. If there is a statistically significant result, then it means that the two populations are unequal (or different).

```
grouped <- data.frame(cbind(data$Analytical_K, data$Experimental_K))
stacked <- stack(grouped)
summary(aov(values ~ ind, data = stacked))

##             Df Sum Sq Mean Sq F value Pr(>F)
## ind          1  0.022  0.0216   0.026  0.875
## Residuals   14 11.777  0.8412

qf(p=0.875, df1=1, df2=14, lower.tail=FALSE)

## [1] 0.02566902
```

(ii)Sum of Squares between and Sum of Squares within

Sum of squares within (SSW):

1. For each subject, compute the difference between its score and its group mean. You thus have to compute each of the group means, and compute the difference between each of the scores and the group mean to which that score belongs
2. Square all these differences

3. Sum the squared differences

Sum of squares between (SSB):

1. For each subject, compute the difference between its group mean and the grand mean. The grand mean is the mean of all NN scores (just sum all scores and divide by the total sample size NN)
2. Square all these differences
3. Sum the squared differences

Sum of Squares(Between) = 0.022
Sum of Squares(Within) = 11.777

(ii) Degrees of Freedom (Between, Within and total)

Degrees of freedom of an estimate is the number of independent pieces of information that went into calculating the estimate. It's not quite the same as the number of items in the sample. In order to get the df for the estimate, you have to subtract 1 from the number of items.

Degrees of Freedom(Between) = 1
Degrees of Freedom(Within) = 14

Degrees of Freedom(Total) = 15

(iii) Mean Squares (Between and Within)

The Mean Sum of Squares between the groups, denoted MSB, is calculated by dividing the Sum of Squares between the groups by the between group degrees of freedom. That is, MSB = SS(Between)/(m−1).

Mean Squares(Between) = 0.0216
Mean Squares(Within) = 0.8412

(iv) F_Statistic and F Critical

The value you calculate from data is called the F Statistic. The F-critical value is a specific value you compare your f-value to. In general, if your calculated F value in a test is larger than your F critical value, then the null hypothesis can be rejected.

F-Statistic = 0.026

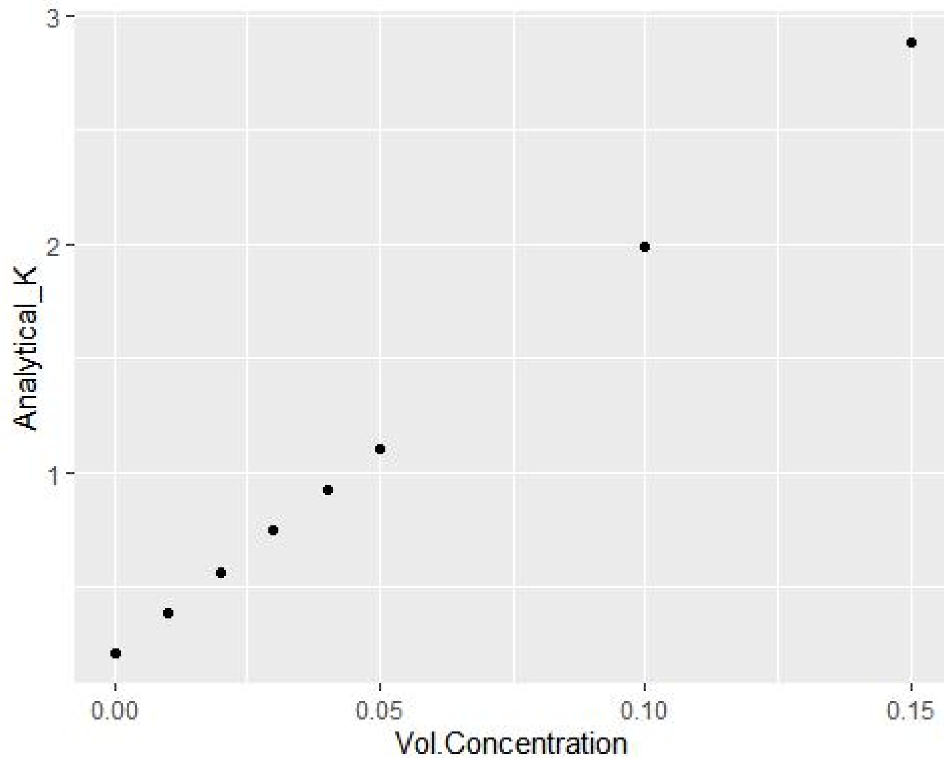
```
fc = qf(p=0.875, df1=1, df2=14, lower.tail=FALSE)
print(paste("F-Crtical:", fc))

## [1] "F-Crtical: 0.0256690170082705"
```

(vi)Check, Is Null hypothesize accepted?

As F-Statistic is larger than F-Critical, **NULL Hypothesis can be rejected.**

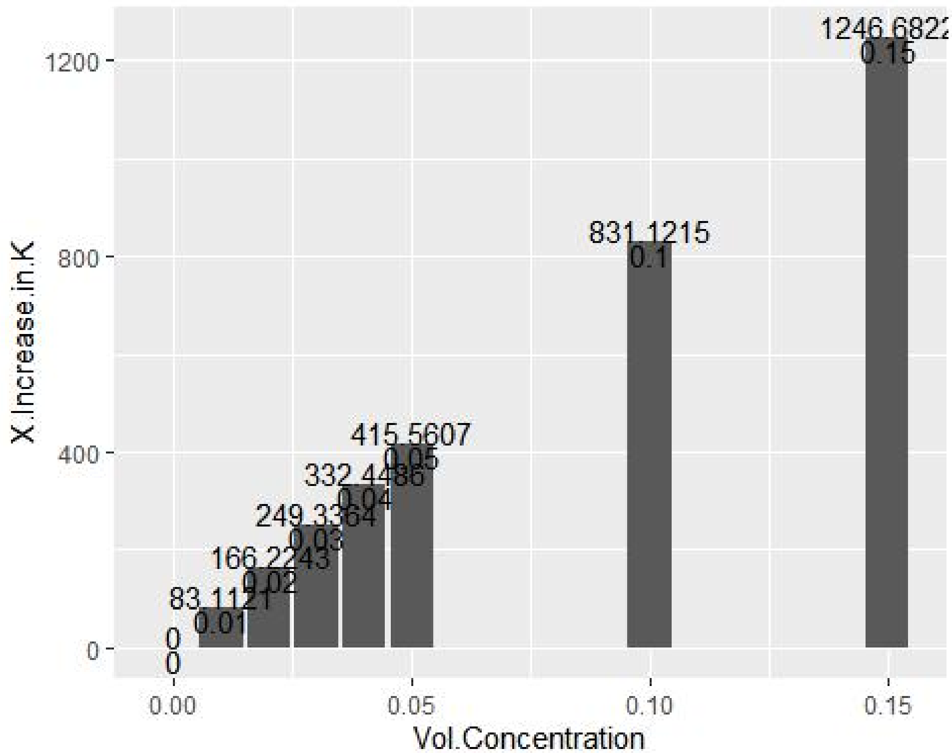
(vii)Plot Vol. Concentration VS Thermal Conductivity

```
ggplot(data,aes(x=Vol.Concentration, y=Analytical_K))+geom_point()
```



(viii)Plot Vol Concentration Vs % increase in Thermal conductivity –(In bar chart)

```
ggplot(data,aes(x=Vol.Concentration, y=X.Increase.in.K)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=X.Increase.in.K), vjust=0) +
  geom_text(aes(label=Vol.Concentration), vjust=1)
```



(ix)Error(Measured Value-Analytical Value)

```
model <- aov(values ~ ind, data = stacked)
sse <- sum((fitted(model) - stacked$values)^2)
print(paste("Sum of Squares Error:", sse))

## [1] "Sum of Squares Error: 11.777023828"

ssr <- sum((fitted(model) - mean(stacked$values))^2)
print(paste("Sum of Squares Regression:", ssr))

## [1] "Sum of Squares Regression: 0.02157961"
```

Conclusions

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic, allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

Here there is no much difference. So a both groups(Analytical & Experimental) values are almost similar.